Massive data sets

**Examples**
- Internet traffic logs
- Financial data
- etc.

**Algorithms**
- Want nearly linear time or less
- Usually at the cost of a randomized approximation
Regression analysis

Regression

- Statistical method to study dependencies between variables in the presence of noise.
Regression analysis

*Linear Regression*

- Statistical method to study **linear** dependencies between variables in the presence of noise.
Regression analysis

*Linear Regression*

- Statistical method to study linear dependencies between variables in the presence of noise.

*Example*

- Ohm's law $V = R \cdot I$
Regression analysis

**Linear Regression**
- Statistical method to study linear dependencies between variables in the presence of noise.

**Example**
- Ohm's law $V = R \cdot I$
- Find linear function that best fits the data

![Example Regression](image-url)
Regression analysis

**Linear Regression**
- Statistical method to study linear dependencies between variables in the presence of noise.

**Standard Setting**
- One measured variable $b$
- A set of predictor variables $a_1, \ldots, a_d$
- Assumption:
  $$ b = x_0 + a_1 x_1 + \ldots + a_d x_d + \varepsilon $$
- $\varepsilon$ is assumed to be noise and the $x_i$ are model parameters we want to learn
- Can assume $x_0 = 0$
- Now consider $n$ observations of $b$
Regression analysis

Matrix form

Input: \( n \times d \)-matrix \( A \) and a vector \( b=(b_1, \ldots, b_n) \)
\( n \) is the number of observations; \( d \) is the number of predictor variables

Output: \( x^* \) so that \( Ax^* \) and \( b \) are close

- Consider the over-constrained case, when \( n \gg d \)
- Can assume that \( A \) has full column rank
Regression analysis

*Least Squares Method*

- Find $x^*$ that minimizes $|Ax-b|^2 = \sum (b_i - \langle A_{i*}, x \rangle)^2$
- $A_{i*}$ is the $i$-th row of $A$
- Certain desirable statistical properties
Regression analysis

Geometry of regression

- We want to find an $x$ that minimizes $|Ax-b|_2$
- The product $Ax$ can be written as

$$A*x_1 + A*x_2 + \ldots + A*x_d$$

where $A*_i$ is the $i$-th column of $A$

- This is a linear $d$-dimensional subspace
- The problem is equivalent to computing the point of the column space of $A$ nearest to $b$ in $l_2$-norm
Regression analysis

Solving least squares regression via the normal equations

- How to find the solution $x$ to $\min_x |Ax-b|_2$?

- Equivalent problem: $\min_x |Ax-b|_2^2$
  - Write $b = Ax' + b'$, where $b'$ orthogonal to columns of $A$
  - Cost is $|A(x-x')|_2^2 + |b'|_2^2$ by Pythagorean theorem
  - Optimal solution $x$ if and only if $A^T(Ax-b) = A^T(Ax-Ax') = 0$
  - Normal Equation: $A^TAx = A^Tb$ for any optimal $x$
  - $x = (A^TA)^{-1} A^T b$

- If the columns of $A$ are not linearly independent, the Moore-Penrose pseudoinverse gives a minimum norm solution $x$
Moore-Penrose Pseudoinverse

**Singular Value Decomposition (SVD)**
Any matrix \( A = U \cdot \Sigma \cdot V^T \)
- \( U \) has orthonormal columns
- \( \Sigma \) is diagonal with non-increasing non-negative entries down the diagonal
- \( V^T \) has orthonormal rows

- Pseudoinverse \( A^- = V \Sigma^{-1} U^T \)
  - Where \( \Sigma^{-1} \) is a diagonal matrix with i-th diagonal entry equal to \( 1/\Sigma_{ii} \) if \( \Sigma_{ii} > 0 \) and is 0 otherwise

- \( \min_x \| Ax - b \|_2^2 \) not unique when columns of \( A \) are linearly independent, but \( x = A^- b \) has minimum norm
Moore-Penrose Pseudoinverse

• Any optimal solution $x$ has the form $A^{-}b + (I - V'V'^{T})z$, where $V'$ corresponds to the rows $i$ of $V^{T}$ for which $\Sigma_{i,i} > 0$

• **Why?**

• Because $A(I - V'V'^{T})z = 0$, so $A^{-}b + (I - V'V'^{T})z$ is a solution. This is a $d$-rank($A$) dimensional affine space so it spans all optimal solutions.

• Since $A^{-}b$ is in column span of $V'$, by Pythagorean theorem, $|A^{-}b + (I - V'V'^{T})z|_{2}^{2} = |A^{-}b|_{2}^{2} + |(I - V'V'^{T})z|_{2}^{2} \geq |A^{-}b|_{2}^{2}$
Time Complexity

Solving least squares regression via the normal equations

- Need to compute $x = A\cdot b$

- Naively this takes $nd^2$ time

- Can do $nd^{1.376}$ using fast matrix multiplication

- But we want much better running time!
Sketching to solve least squares regression

- How to find an approximate solution $x$ to $\min_x |Ax-b|_2$?

- **Goal:** output $x'$ for which $|Ax'-b|_2 \leq (1+\epsilon) \min_x |Ax-b|_2$ with high probability

- Draw $S$ from a $k \times n$ random family of matrices, for a value $k << n$

- Compute $S^*A$ and $S^*b$

- Output the solution $x'$ to $\min_{x'} |(SA)x-(Sb)|_2$
  - $x' = (SA)^{-1}Sb$
How to choose the right sketching matrix $S$?

- Recall: output the solution $x'$ to $\min_{x'} |(SA)x-(Sb)|_2$
- Lots of matrices work
- $S$ is $d/\varepsilon^2 \times n$ matrix of i.i.d. Normal random variables
- To see why this works, we introduce the notion of a subspace embedding
Subspace Embeddings

• Let $k = O(d/\varepsilon^2)$
• Let $S$ be a $k \times n$ matrix of i.i.d. normal $N(0,1/k)$ random variables
• For any fixed $d$-dimensional subspace, i.e., the column space of an $n \times d$ matrix $A$
  – W.h.p., for all $x$ in $\mathbb{R}^d$, $|SAx|_2 = (1\pm\varepsilon)|Ax|_2$
• Entire column space of $A$ is preserved

Why is this true?
Subspace Embeddings – A Proof

• Want to show $|SAx|^2 = (1 \pm \varepsilon)|Ax|^2$ for all $x$

• Can assume columns of $A$ are orthonormal (since we prove this for all $x$)

• Claim: $SA$ is a $k \times d$ matrix of i.i.d. $N(0,1/k)$ random variables

  – First property: for two independent random variables $X$ and $Y$, with $X$ drawn from $N(0,a^2)$ and $Y$ drawn from $N(0,b^2)$, we have $X+Y$ is drawn from $N(0, a^2 + b^2)$
X+Y is drawn from \( \text{N}(0, a^2 + b^2) \)

- Probability density function \( f_z \) of \( Z = X+Y \) is convolution of probability density functions \( f_x \) and \( f_y \)

\[
f_z(z) = \int f_y(z - x)f_x(x) \, dx
\]

- \( f_x(x) = \frac{1}{a(2\pi)^{\frac{5}{2}}} e^{-x^2/2a^2} \), \( f_y(y) = \frac{1}{b(2\pi)^{\frac{5}{2}}} e^{-y^2/2b^2} \)

- \( f_z(z) = \int \frac{1}{a(2\pi)^{\frac{5}{2}}} e^{-(z-x)^2/2a^2} \frac{1}{b(2\pi)^{\frac{5}{2}}} e^{-x^2/2b^2} \, dx
\]

\[
= \frac{1}{(2\pi)^{\frac{5}{2}}(a^2+b^2)^{\frac{5}{2}}} e^{-z^2/2(a^2+b^2)} \int \frac{(a^2+b^2)^{\frac{5}{2}}}{(2\pi)^{\frac{5}{2}}ab} e^{-2\left(\frac{(ab)^2}{a^2+b^2}\right)} \, dx
\]
X+Y is drawn from \( N(0, a^2 + b^2) \)

Calculation: 
\[
\int e^{-\frac{(z-x)^2}{2a^2}} \frac{x^2}{2b^2} = e
\]

Density of Gaussian distribution: 
\[
\int \frac{(a^2+b^2)^5}{(2\pi)^5 ab} e^{-\frac{(x-\frac{b^2 z}{a^2+b^2})^2}{2\left(\frac{(ab)^2}{a^2+b^2}\right)}} dx = 1
\]
Rotational Invariance

• Second property: if \( u, v \) are vectors with \( \langle u, v \rangle = 0 \), then \( \langle g, u \rangle \) and \( \langle g, v \rangle \) are independent, where \( g \) is a vector of i.i.d. \( N(0,1/k) \) random variables

• Why?

• If \( g \) is an \( n \)-dimensional vector of i.i.d. \( N(0,1) \) random variables, and \( R \) is a fixed matrix, then the probability density function of \( Rg \) is

\[
f(x) = \frac{1}{\det(RR^T)(2\pi)^{d/2}} e^{-\frac{x^T(RR^T)^{-1}x}{2}}
\]

- \( RR^T \) is the covariance matrix
- For a rotation matrix \( R \), the distribution of \( Rg \) and of \( g \) are the same
Orthogonal Implies Independent

• Want to show: if $u, v$ are vectors with $\langle u, v \rangle = 0$, then $\langle g, u \rangle$ and $\langle g, v \rangle$ are independent, where $g$ is a vector of i.i.d. $N(0, 1/k)$ random variables.

• Choose a rotation $R$ which sends $u$ to $\alpha e_1$, and sends $v$ to $\beta e_2$.

$$\langle g, u \rangle = \langle gR, R^Tu \rangle = \langle h, \alpha e_1 \rangle = \alpha h_1$$

$$\langle g, v \rangle = \langle gR, R^Tv \rangle = \langle h, \beta e_2 \rangle = \beta h_2$$

where $h$ is a vector of i.i.d. $N(0, 1/k)$ random variables.

• Then $h_1$ and $h_2$ are independent by definition.
Where were we?

- **Claim:** SA is a k x d matrix of i.i.d. N(0,1/k) random variables

- **Proof:** The rows of SA are independent
  - Each row is: \(< g, A_1 >, < g, A_2 >, ..., < g, A_d >\)
  - First property implies the entries in each row are N(0,1/k) since the columns $A_i$ have unit norm
  - Since the columns $A_i$ are orthonormal, the entries in a row are independent by our second property
Back to Subspace Embeddings

- Want to show $|SAx|_2 = (1 \pm \epsilon)|Ax|_2$ for all $x$
- Can assume columns of $A$ are orthonormal
- Can also assume $x$ is a unit vector
- $SA$ is a $k \times d$ matrix of i.i.d. $N(0, 1/k)$ random variables

- Consider any fixed unit vector $x \in R^d$
- $|SAx|_2^2 = \sum_{i \in [k]} < g_i, x >^2$, where $g_i$ is $i$-th row of $SA$
- Each $< g_i, x >^2$ is distributed as $N\left(0, \frac{1}{k}\right)^2$
- $E[< g_i, x >^2] = 1/k$, and so $E[|SAx|_2^2] = 1$

How concentrated is $|SAx|_2^2$ about its expectation?
Johnson-Lindenstrauss Theorem

• Suppose $h_1, \ldots, h_k$ are i.i.d. $N(0,1)$ random variables
• Then $G = \sum_i h_i^2$ is a $\chi^2$-random variable
• Apply known tail bounds to $G$: 
  – (Upper) $\Pr[G \geq k + 2(kx)^5 + 2x] \leq e^{-x}$
  – (Lower) $\Pr[G \leq k - 2(kx)^5] \leq e^{-x}$
• If $x = \frac{\epsilon^2 k}{16}$, then $\Pr[G \in k(1 \pm \epsilon)] \geq 1 - 2e^{-\epsilon^2 k/16}$
• If $k = \Theta(\epsilon^{-2} \log(\frac{1}{\delta}))$, this probability is $1-\delta$

• $\Pr[|S\mathbf{x}|^2_2 \in (1 \pm \epsilon)] \geq 1 - 2^{-\Theta(d)}$

*This only holds for a fixed $x$, how to argue for all $x$?*
Net for Sphere

• Consider the sphere $S^{d-1}$

• Subset $N$ is a $\gamma$-net if for all $x \in S^{d-1}$, there is a $y \in N$, such that $|x - y|_2 \leq \gamma$

• Greedy construction of $N$
  – While there is a point $x \in S^{d-1}$ of distance larger than $\gamma$ from every point in $N$, include $x$ in $N$

• The sphere of radius $\gamma/2$ around every point in $N$ is contained in the sphere of radius $1+\gamma/2$ around $0^d$

• Further, all such spheres are disjoint

• Ratio of volume of $d$-dimensional sphere of radius $1+\gamma/2$ to dimensional sphere of radius $\gamma$ is $(1 + \gamma/2)^d/(\gamma/2)^d$, so $|N| \leq (1 + \gamma/2)^d/(\gamma/2)^d$
Net for Subspace

- Let \( M = \{Ax \mid x \text{ in } N\} \), so \( |M| \leq (1 + \gamma/2)^d/\gamma^d \)

- Claim: For every \( x \) in \( S^{d-1} \), there is a \( y \) in \( M \) for which \( |Ax - y|_2 \leq \gamma \)

- Proof: Let \( x' \) in \( S^{d-1} \) be such that \( |x - x'|_2 \leq \gamma \). Then \( |Ax - Ax'|_2 = |x - x'|_2 \leq \gamma \), using that the columns of \( A \) are orthonormal. Set \( y = Ax' \)
Net Argument

• For a fixed unit $x$, $\Pr[|SAx|^2_2 \in (1 \pm \epsilon)] \geq 1 - 2^{-\Theta(d)}$
• For a fixed pair of unit $x$, $x'$, $|SAx|^2_2$, $|SAx'|^2_2$, $|SA(x - x')|^2_2$ are all $1 \pm \epsilon$ with probability $1 - 2^{-\Theta(d)}$
• $|SA(x - x')|^2_2 = |SAx|^2_2 + |SAx'|^2_2 - 2 < SAx, SAx' >$
• $|A(x - x')|^2_2 = |Ax|^2_2 + |Ax'|^2_2 - 2 < Ax, Ax' >$
  - So $\Pr[< Ax, Ax' > = < SAx, SAx' > \pm 0(\epsilon)] = 1 - 2^{-\Theta(d)}$
• Choose a $\frac{1}{2}$-net $M = \{Ax | x \in N\}$ of size $5^d$
• By a union bound, for all pairs $y$, $y'$ in $M$,
  $< y, y' > = < Sy, Sy' > \pm 0(\epsilon)$
• Condition on this event
• By linearity, if this holds for $y$, $y'$ in $M$, for $\alpha y$, $\beta y'$ we have
  $< \alpha y, \beta y' > = \alpha \beta < Sy, Sy' > \pm 0(\epsilon \alpha \beta)$
Finishing the Net Argument

- Let $y = Ax$ for an arbitrary $x \in S^{d-1}$
- Let $y_1 \in M$ be such that $|y - y_1|_2 \leq \gamma$
- Let $\alpha$ be such that $|\alpha(y - y_1)|_2 = 1$
  - $\alpha \geq 1/\gamma$ (could be infinite)
- Let $y'_2 \in M$ be such that $|\alpha(y - y_1) - y'_2|_2 \leq \gamma$
- Then $\left| y - y_1 - \frac{y'_2}{\alpha} \right|_2 \leq \frac{\gamma}{\alpha} \leq \gamma^2$
- Set $y_2 = \frac{y'_2}{\alpha}$. Repeat, obtaining $y_1, y_2, y_3, ...$ such that for all integers $i$,
  $$|y - y_1 - y_2 - ... - y_i|_2 \leq \gamma^i$$
- Implies $|y_i|_2 \leq \gamma^{i-1} + \gamma^i \leq 2\gamma^{i-1}$
Finishing the Net Argument

- Have $y_1, y_2, y_3, \ldots$ such that $y = \sum_i y_i$ and $|y_i|_2 \leq 2\gamma^{i-1}$

- $|Sy|_2^2 = |S \sum_i y_i|_2^2$
  $= \sum_i |Sy_i|_2^2 + 2 \sum_{i,j} < Sy_i, Sy_j >$
  $= \sum_i |y_i|_2^2 + 2 \sum_{i,j} < y_i, y_j > \pm O(\varepsilon) \sum_{i,j} |y_i|_2 |y_j|_2$
  $= |\sum_i y_i|_2^2 \pm O(\varepsilon)$
  $= |y|_2^2 \pm O(\varepsilon)$
  $= 1 \pm O(\varepsilon)$

- Since this held for an arbitrary $y = Ax$ for unit $x$, by linearity it follows that for all $x$, $|SAx|_2 = (1\pm\varepsilon)|Ax|_2$
Back to Regression

• We showed that S is a subspace embedding, that is, simultaneously for all \( x \),
  \[ |SAx|_2 = (1 \pm \varepsilon)|Ax|_2 \]

What does this have to do with regression?
Subspace Embeddings for Regression

- Want $x$ so that $|Ax-b|_2 \leq (1+\varepsilon) \min_y |Ay-b|_2$
- Consider subspace $L$ spanned by columns of $A$ together with $b$
- Then for all $y$ in $L$, $|Sy|_2 = (1 \pm \varepsilon) |y|_2$
- Hence, $|S(Ax-b)|_2 = (1 \pm \varepsilon) |Ax-b|_2$ for all $x$
- Solve $\arg\min_y |(SA)y - (Sb)|_2$
- Given $SA$, $Sb$, can solve in poly($d/\varepsilon$) time

Only problem is computing $SA$ takes $O(nd^2)$ time
How to choose the right sketching matrix $S$? $[S]$

- $S$ is a Subsampled Randomized Hadamard Transform
  - $S = P^*H^*D$

- $D$ is a diagonal matrix with $+1$, $-1$ on diagonals

- $H$ is the Hadamard transform

- $P$ just chooses a random (small) subset of rows of $H^*D$

- $S^*A$ can be computed in $O(nd \log n)$ time

Why does it work?
Why does this work?

- We can again assume columns of A are orthonormal

- It suffices to show \( |S Ax|_2^2 = |PHDAx|_2^2 = 1 \pm \epsilon \) for all \( x \)

- HD is a rotation matrix, so \( |HDAx|_2^2 = |Ax|_2^2 = 1 \) for any \( x \)
  - Notation: let \( y = Ax \)

- Flattening Lemma: For any fixed \( y \),
  \[
  \Pr \left[ |HDy|_\infty \geq C \frac{\log^5 nd/\delta}{n^5} \right] \leq \frac{\delta}{2d}
  \]
Proving the Flattening Lemma

- **Flattening Lemma**: \( \Pr \left[ |HDy|_\infty \geq C \frac{\log^5 nd}{n^5} \right] \leq \frac{\delta}{2d} \)

- Let \( C > 0 \) be a constant. We will show for a fixed \( i \) in \([n]\),

\[
\Pr \left[ |(HDy)_i| \geq C \frac{\log^5 nd}{n^5} \right] \leq \frac{\delta}{2nd}
\]

- If we show this, we can apply a union bound over all \( i \)

\[
|(HDy)_i| = \sum_j H_{i,j} D_{j,i} y_j
\]

- (Azuma-Hoeffding) \( \Pr[|\sum_j Z_j| > t] \leq 2e^{-\frac{t^2}{2 \sum_j \beta_j^2}} \), where \( |Z_j| \leq \beta_j \) with probability 1
  - \( Z_j = H_{i,j} D_{j,i} y_j \) has 0 mean
  - \( |Z_j| \leq \frac{|y_j|}{n^5} = \beta_j \) with probability 1
  - \( \sum_j \beta_j^2 = \frac{1}{n} \)

\[
\Pr \left[ |\sum Z_j| > C \frac{\log^5 (nd)}{n^5} \right] \leq 2e^{-\frac{c^2 \log(nd)}{2}} \leq \frac{\delta}{2nd}
\]
Consequence of the Flattening Lemma

- Recall columns of $A$ are orthonormal
- HDA has orthonormal columns
- Flattening Lemma implies $|HDAe_i|_\infty \leq C \frac{\log^5 nd/\delta}{n^5}$ with probability $1 - \frac{\delta}{2d}$ for a fixed $i \in [d]$
- With probability $1 - \frac{\delta}{2}$, $|e_jHDAe_i| \leq C \frac{\log^5 nd/\delta}{n^5}$ for all $i,j$
- Given this, $|e_jHDA|_2 \leq C \frac{d^5 \log^5 nd/\delta}{n^5}$ for all $j$

(Can be optimized further)
Matrix Chernoff Bound

- Let $X_1, \ldots, X_s$ be independent copies of a symmetric random matrix $X \in \mathbb{R}^{d \times d}$ with $E[X] = 0$, $|X|_2 \leq \gamma$, and $|E[X^TX]|_2 \leq \sigma^2$. Let $W = \frac{1}{s} \sum_{i \in [s]} X_i$. For any $\epsilon > 0$,

$$\Pr[|W|_2 > \epsilon] \leq 2d \cdot e^{-s\epsilon^2/(\sigma^2 + \gamma^2/3)}$$

(here $|W|_2 = \sup |Wx|_2/|x|_2$)

- Let $V = \text{HDA}$, and recall $V$ has orthonormal columns

- Suppose $P$ in the $S = \text{PHD}$ definition samples uniformly with replacement. If row $i$ is sampled in the $j$-th sample, then $P_{j,i} = n$, and is 0 otherwise

- Let $Y_i$ be the $i$-th sampled row of $V = \text{HDA}$

- Let $X_i = I_d - n \cdot Y_i^TY_i$
  - $E[X_i] = I_d - n \cdot \sum_j \left(\frac{1}{n}\right) V_j^TV_j = I_d - V^TV = 0^d$
  - $|X_i|_2 \leq |I_d|_2 + n \cdot \max |e_j\text{HDA}|^2 = 1 + n \cdot C^2 \log \left(\frac{nd}{\delta}\right) \cdot \frac{d}{n} = \Theta(d \log \left(\frac{nd}{\delta}\right))$
Matrix Chernoff Bound

- Recall: let $Y_i$ be the $i$-th sampled row of $V = HDA$
- Let $X_i = I_d - n \cdot Y_i^T Y_i$
- $E[X^T X + I_d] = I_d + I_d - 2n E[Y_i^T Y_i] + n^2 E[Y_i^T Y_i Y_i^T Y_i]$
  \[= 2I_d - 2I_d + n^2 \sum_i \left( \frac{1}{n} \right) \cdot v_i^T v_i v_i^T v_i = n \sum_i v_i^T v_i \cdot |v_i|^2 \]
- Define $Z = n \sum_i v_i^T v_i \cdot C^2 \log \left( \frac{nd}{\delta} \right) \cdot \frac{d}{n} = C^2 d \log \left( \frac{nd}{\delta} \right) I_d$
- Note that $E[X^T X + I_d]$ and $Z$ are real symmetric, with non-negative eigenvalues
- Claim: for all vectors $y$, we have: $y^T E[X^T X + I_d] y \leq y^T Z y$
- Proof: $y^T E[X^T X + I_d] y = n \sum_i y^T v_i^T v_i y |v_i|^2 = n \sum_i < v_i, y >^2 |v_i|^2$ and
  \[y^T Z y = n \sum_i y^T v_i^T v_i y \cdot C^2 \log \left( \frac{nd}{\delta} \right) \cdot \frac{d}{n} = d \sum_i < v_i, y >^2 C^2 \log \left( \frac{nd}{\delta} \right)\]
- Hence, $|E[X^T X]|_2 \leq |E[X^T X] + I_d|_2 + |I_d|_2 = |E[X^T X + I_d]|_2 + 1$
  \[\leq |Z|_2 + 1 \leq C^2 d \log \left( \frac{nd}{\delta} \right) + 1\]
- Hence, $|E[X^T X]|_2 = O \left( d \log \left( \frac{nd}{\delta} \right) \right)$
Matrix Chernoff Bound

- Hence, $|E[X^TX]|_2 = O\left(d \log \left(\frac{nd}{\delta}\right)\right)$

- Recall: (Matrix Chernoff) Let $X_1, \ldots, X_s$ be independent copies of a symmetric random matrix $X \in \mathbb{R}^{d \times d}$ with $E[X] = 0$, $|X|_2 \leq \gamma$, and $|E[X^TX]|_2 \leq \sigma^2$. Let $W = \frac{1}{s} \sum_{i \in [s]} X_i$. For any $\epsilon > 0$, $\Pr[|W|_2 > \epsilon] \leq 2d \cdot e^{-s\epsilon^2/((\sigma^2 + \frac{\gamma^2}{3})}$

$$\Pr\left[|I_d - (PHDA)^T(\text{PHDA})|_2 > \epsilon\right] \leq 2d \cdot e^{-s\epsilon^2/(\Theta(d \log \left(\frac{nd}{\delta}\right))}$$

- Set $s = d \log \left(\frac{nd}{\delta}\right) \log \left(\frac{d}{\delta}\right)$, to make this probability less than $\frac{\delta}{2}$
SRHT Wrapup

- Have shown \(|I_d - (PHDA)^T(PHDA)|_2 < \epsilon\) using Matrix Chernoff Bound and with \(s = d \log\left(\frac{nd}{\delta}\right) \frac{\log(d)}{\epsilon^2}\) samples.

- Implies for every unit vector \(x\),
  \[|1 - |PHDAx|^2_2| = |x^T x - x^T (PHDA)^T (PHDA)x| < \epsilon\],
  so \(|PHDAx|^2_2 \in 1 \pm \epsilon\) for all unit vectors \(x\).

- Considering the column span of \(A\) adjoined with \(b\), we can again solve the regression problem.

- The time for regression is now only \(O(nd \log n) + \text{poly}\left(\frac{d \log(n)}{\epsilon}\right)\). Nearly optimal in matrix dimensions \((n >> d)\).
Faster Subspace Embeddings S [CW,MM,NN]

- CountSketch matrix

- Define $k \times n$ matrix $S$, for $k = O(d^2/\varepsilon^2)$

- $S$ is really sparse: single randomly chosen non-zero entry per column

\[
\begin{bmatrix}
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 1 & 0 & -1 & 0 \\
0 & -1 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]

- $\text{nnz}(A)$ is number of non-zero entries of $A$

Can compute $S \cdot A$ in $\text{nnz}(A)$ time!
Simple Proof [Nguyen]

- Can assume columns of A are orthonormal

- Suffices to show $|SAx|_2 = 1 \pm \epsilon$ for all unit $x$
  - For regression, apply $S$ to $[A, b]$

- $SA$ is a $2d^2/\epsilon^2 \times d$ matrix

- Suffices to show $|A^T S^T SA - I|_2 \leq |A^T S^T SA - I|_F \leq \epsilon$

- Matrix product result shown below:
  $$\Pr[|CS^TSD - CD|_F^2 \leq \frac{6}{(\delta(\# \text{ rows of } S))} \cdot |C|_F^2 |D|_F^2] \geq 1 - \delta$$

- Set $C = A^T$ and $D = A$.

- Then $|A|_F^2 = d$ and $(\# \text{ rows of } S) = 6d^2/(\delta \epsilon^2)$
Matrix Product Result [Kane, Nelson]

- Show: \( \Pr[|CS^TSD - CD|_F^2 \leq [6/(\delta(\# \text{ rows of } S))] \cdot |C|_F^2 |D|_F^2] \geq 1 - \delta \)

- (JL Property) A distribution on matrices \( S \in \mathbb{R}^{k \times n} \) has the \((\epsilon, \delta, \ell)\)-JL moment property if for all \( x \in \mathbb{R}^n \) with \(|x|_2 = 1\),
  \[ E_S||Sx|_2^2 - 1|^{\ell} \leq \epsilon^{\ell} \cdot \delta \]

- (From vectors to matrices) For \( \epsilon, \delta \in \left(0, \frac{1}{2}\right)\), let \( D \) be a distribution on matrices \( S \) with \( k \) rows and \( n \) columns that satisfies the \((\epsilon, \delta, \ell)\)-JL moment property for some \( \ell \geq 2 \). Then for \( A, B \) matrices with \( n \) rows,
  \[ \Pr_S \left[ |A^T S^T S B - A^T B|_F \geq 3 \epsilon |A|_F |B|_F \right] \leq \delta \]
(From vectors to matrices) For $\epsilon, \delta \in \left(0, \frac{1}{2}\right)$, let $D$ be a distribution on matrices $S$ with $k$ rows and $n$ columns that satisfies the $(\epsilon, \delta, \ell)$-JL moment property for some $\ell \geq 2$. Then for $A, B$ matrices with $n$ rows,

$$\Pr_S \left[ |A^T S^T S B - A^T B|_F \geq 3 \epsilon |A|_F |B|_F \right] \leq \delta$$

**Proof:** For a random scalar $X$, let $|X|_p = (E|X|^p)^{1/p}$

- Sometimes consider $X = |T|_F$ for a random matrix $T$
- $| |T|_F |_p = (E[|T|_F^p])^{1/p}$

Can show $|. |_p$ is a norm if $p \geq 1$

- Minkowski’s Inequality: $|X + Y|_p \leq |X|_p + |Y|_p$

For unit vectors $x, y$, we will bound $|\langle Sx, Sy \rangle - \langle x, y \rangle|_\ell$
Minkowski’s Inequality

- Minkowski’s Inequality: $|X + Y|_p \leq |X|_p + |Y|_p$
- Proof:
  - If $|X|_p$, $|Y|_p$ are finite, then so is $|X + Y|_p$. Why?
  - $f(x) = x^p$ is convex for $p \geq 1$, so for any fixed $x$, $y$:
    
    $|.5x + .5y|^p \leq |.5|x| + .5|y||^p \leq .5|x|^p + .5|y|^p$, so
    
    $|x + y|^p \leq 2^{p-1}(|x|^p + |y|^p)$
  - So, $E[|X + Y|_p^p] \leq E[2^{p-1}(|X|_p^p + |Y|_p^p)]$

- $|X + Y|_p^p = \int |x + y|^p d\mu$
  
  $= \int |x + y| \cdot |x + y|^{p-1} d\mu$
  
  $\leq \int (|x| + |y|)|x + y|^{p-1} d\mu$
  
  $= \int |x||x + y|^{p-1} d\mu + \int |y||x + y|^{p-1} d\mu$

  $\leq \left(\left(\int |x|^p d\mu\right)^{\frac{1}{p}} + \left(\int |y|^p d\mu\right)^{\frac{1}{p}}\right) \left(\int |x + y|^{(p-1)\left(\frac{p}{p-1}\right)} d\mu\right)^{\frac{p-1}{p}}$

  $= (|X|_p + |Y|_p)|X + Y|_p^{p-1}$
From Vectors to Matrices

- For unit vectors \( x, y \), \( \| \langle Sx, Sy \rangle - \langle x, y \rangle \|_\ell \)

\[
= \frac{1}{2} \left( (|Sx|_2^2 - 1) + (|Sy|_2^2 - 1) - (|S(x - y)|_2^2 - |x - y|_2^2) \right) \|_\ell \\
\leq \frac{1}{2} \left( (|Sx|_2^2 - 1) + (|Sy|_2^2 - 1) + (|S(x - y)|_2^2 - |x - y|_2^2) \right) \|_\ell \\
\leq \frac{1}{2} \left( \epsilon \cdot \delta + \epsilon \cdot \delta + |x - y|_2^2 \epsilon \cdot \delta \right) \\
\leq 3 \epsilon \cdot \delta \|_\ell
\]

- By linearity, for arbitrary \( x, y \), \( \frac{\| \langle Sx, Sy \rangle - \langle x, y \rangle \|_\ell}{|x|_2|y|_2} \leq 3 \epsilon \cdot \delta \|_\ell \)

- Suppose \( A \) has \( d \) columns and \( B \) has \( e \) columns. Let the columns of \( A \) be \( A_1, ..., A_d \) and the columns of \( B \) be \( B_1, ..., B_e \)

- Define \( X_{i,j} = \frac{1}{|A_i|_2|B_j|_2} \cdot (\langle SA_i, SB_j \rangle - \langle A_i, B_j \rangle) \)

- \( \| A^T S^T S B - A^T B \|_F^2 = \sum_i \sum_j |A_i|_2^2 \cdot |B_j|_2^2 X_{i,j}^2 \)
From Vectors to Matrices

- Have shown: for arbitrary \( x, y \), \( \frac{|\langle Sx, Sy \rangle - \langle x, y \rangle|_\ell}{|x|_2 |y|_2} \leq 3 \epsilon \cdot \delta_\ell \)

- For \( X_{i,j} = \frac{1}{|A_i|_2 |B_j|_2} \cdot (\langle SA_i, SB_j \rangle - \langle A_i, B_j \rangle) \): \( |A^T S^T S B - A^T B|_F^2 = \Sigma_i \Sigma_j |A_i|_2^2 \cdot |B_j|_2^2 X_{i,j}^2 \)

- \( |A^T S^T S B - A^T B|_F^2 |\ell/2 = |\Sigma_i \Sigma_j |A_i|_2^2 \cdot |B_j|_2^2 X_{i,j}^2|_\ell/2 \)
  \[ \leq \Sigma_i \Sigma_j |A_i|_2^2 \cdot |B_j|_2^2 |X_{i,j}|_\ell \]
  \[ = \Sigma_i \Sigma_j |A_i|_2^2 \cdot |B_j|_2^2 |X_{i,j}|^2_\ell \]
  \[ \leq (3\epsilon \delta_\ell^2)^2 \Sigma_i \Sigma_j |A_i|_2^2 |B_j|_2^2 \]
  \[ = (3\epsilon \delta_\ell^2)^2 |A|_F^2 |B|_F^2 \]

- Since \( E \left[ |A^T S^T S B - A^T B|_F^\ell \right] = \left[ |A^T S^T S B - A^T B|_F^2 \right]^{\ell/2} \), by Markov’s inequality,

- \( \text{Pr} \left[ |A^T S^T S B - A^T B|_F > 3\epsilon |A|_F |B|_F \right] \leq \left( \frac{1}{3\epsilon |A|_F |B|_F} \right)^\ell \text{E}[|A^T S^T S B - A^T B|_F^\ell] \leq \delta \)
Result for Vectors

- **Show:** \( \Pr[|CSTSD – CD|_F^2 \leq \frac{[6/(\delta(\text{# rows of S}))]}{6}|C|_F^2 |D|_F^2] \geq 1 - \delta \)

- *(JL Property)* A distribution on matrices \( S \in \mathbb{R}^{k \times n} \) has the \((\epsilon, \delta, \ell)\)-JL moment property if for all \( x \in \mathbb{R}^n \) with \( |x|_2 = 1 \),
  \[
  \mathbb{E}_S \left( |Sx|_2^\ell - 1 \right)^\ell \leq \epsilon^\ell \cdot \delta
  \]

- *(From vectors to matrices)* For \( \epsilon, \delta \in \left(0, \frac{1}{2}\right) \), let \( D \) be a distribution on matrices \( S \) with \( k \) rows and \( n \) columns that satisfies the \((\epsilon, \delta, \ell)\)-JL moment property for some \( \ell \geq 2 \). Then for \( A, B \) matrices with \( n \) rows,
  \[
  \Pr_{S} \left[ |A^TSTSB – A^TB|_F \geq 3 \epsilon |A|_F |B|_F \right] \leq \delta
  \]

- Just need to show that the CountSketch matrix \( S \) satisfies JL property and bound the number \( k \) of rows
CountSketch Satisfies the JL Property

- **(JL Property)** A distribution on matrices $S \in \mathbb{R}^{k \times n}$ has the $(\epsilon, \delta, \ell)$-JL moment property if for all $x \in \mathbb{R}^n$ with $|x|_2 = 1$,
  $$E_S \left| |Sx|_2^2 - 1 \right|^\ell \leq \epsilon^\ell \cdot \delta$$

- We show this property holds with $\ell = 2$. First, let us consider $\ell = 1$

- For CountSketch matrix $S$, let
  - $h: [n] \rightarrow [k]$ be a 2-wise independent hash function
  - $\sigma: [n] \rightarrow \{-1,1\}$ be a 4-wise independent hash function

- Let $\delta(E) = 1$ if event $E$ holds, and $\delta(E) = 0$ otherwise

- $E[|Sx|_2^2] = \sum_{j \in [k]} E\left[ \left( \sum_{i \in [n]} \delta(h(i) = j) \sigma_i x_i \right)^2 \right]$
  $$= \sum_{j \in [k]} \sum_{i_1, i_2 \in [n]} E[\delta(h(i1) = j) \delta(h(i2) = j) \sigma_{i1} \sigma_{i2} x_{i1} x_{i2}]$$
  $$= \sum_{j \in [k]} \sum_{i \in [n]} E[\delta(h(i) = j)^2] x_i^2$$
  $$= \left( \frac{1}{k} \right) \sum_{j \in [k]} \sum_{i \in [n]} x_i^2 = |x|_2^2$$
CountSketch Satisfies the JL Property

- $E[|Sx|^4] = E[\sum_{j\in[k]} \sum_{j'\in[k]} \left( \sum_{i\in[n]} \delta(h(i) = j)\sigma_i x_i \right)^2 \left( \sum_{i'\in[n]} \delta(h(i') = j')\sigma_{i'} x_{i'} \right)^2] = $
  
  $\sum_{j_1,j_2,i_1,i_2,i_3,i_4} E[\sigma_{i_1} \sigma_{i_2} \sigma_{i_3} \sigma_{i_4} \delta(h(i_1) = j_1) \delta(h(i_2) = j_1) \delta(h(i_3) = j_2) \delta(h(i_4 = j_2))] x_{i_1} x_{i_2} x_{i_3} x_{i_4}$

- We must be able to partition $\{i_1, i_2, i_3, i_4\}$ into equal pairs.

- Suppose $i_1 = i_2 = i_3 = i_4$. Then necessarily $j_1 = j_2$. Obtain $\sum_{j \in[k]} \sum_i x_i^4 = |x|^4$

- Suppose $i_1 = i_2$ and $i_3 = i_4$ but $i_1 \neq i_3$. Then get $\sum_{j_1,j_2,i_1,i_3} \sum_{k=1}^k \frac{1}{k^2} x_{i_1}^2 x_{i_3}^2 = \frac{1}{k} |x|^4 - |x|^4$

- Suppose $i_1 = i_3$ and $i_2 = i_4$ but $i_1 \neq i_2$. Then necessarily $j_1 = j_2$. Obtain $\sum_{j \in[k]} \sum_{i_1,i_2} x_{i_1}^2 x_{i_2}^2 \leq \frac{1}{k} |x|^4$. Obtain same bound if $i_1 = i_4$ and $i_2 = i_3$.

- Hence, $E[|Sx|^4] \in [|x|^4, |x|^4(1 + \frac{2}{k})] = [1, 1 + \frac{2}{k}]$

- So, $E_S||Sx|_2^2 - 1|^2 \leq \left(1 + \frac{2}{k}\right) - 2 + 1 = \frac{2}{k}$. Setting $k = \frac{2}{\epsilon^2 \delta}$ finishes the proof.
(JL Property) A distribution on matrices $S \in \mathbb{R}^{k \times n}$ has the $(\epsilon, \delta, \ell)$-JL moment property if for all $x \in \mathbb{R}^n$ with $|x|_2 = 1$,

$$E_S \left| |Sx|_2^\ell - 1 \right|^\ell \leq \epsilon^\ell \cdot \delta$$

(From vectors to matrices) For $\epsilon, \delta \in \left(0, \frac{1}{2}\right)$, let $D$ be a distribution on matrices $S$ with $k$ rows and $n$ columns that satisfies the $(\epsilon, \delta, \ell)$-JL moment property for some $\ell \geq 2$. Then for $A, B$ matrices with $n$ rows,

$$\Pr \left[ \left| A^T S^T S B - A^T B \right|_F^2 \geq 3 \epsilon^2 |A|_F^2 |B|_F^2 \right] \leq \delta$$

We showed CountSketch has the JL property with $\ell = 2$, and $k = \frac{2}{\epsilon^2 \delta}$

Matrix product result we wanted was:

$$\Pr[|CS^T S D - C D|_F^2 \leq (6/(\delta k)) \cdot |C|_F^2 |D|_F^2] \geq 1 - \delta$$

We are now done with the proof CountSketch is a subspace embedding
Course Outline

- Subspace embeddings and least squares regression
  - Gaussian matrices
  - Subsampled Randomized Hadamard Transform
  - CountSketch
- **Affine embeddings**
  - Application to low rank approximation
- High precision regression
- Leverage score sampling
- Distributed low rank approximation
- L1 Regression
- M-Estimator regression
Affine Embeddings

- Want to solve $\min_X |AX - B|^2_F$, $A$ is tall and thin with $d$ columns, but $B$ has a large number of columns

- Can't directly apply subspace embeddings

- Let's try to show $|SAX - SB|^2_F = (1 \pm \epsilon)|AX - B|^2_F$ for all $X$ and see what properties we need of $S$

- Can assume $A$ has orthonormal columns

- Let $B^* = AX^* - B$, where $X^*$ is the optimum

| $|S(AX - B)|^2_F - |SB^*|^2_F = |SA(X - X^*) + S(AX^* - B)|^2_F - |SB^*|^2_F$ |
|-----------------------------------------------|
| $= |SA(X - X^*)|^2_F + 2\text{tr}[(X - X^*)^T A^T S^T S B^*]$ (use $|C + D|^2_F = |C|^2_F + |D|^2_F + 2\text{tr}(C^T D)$) |
| $\in |SA(X - X^*)|^2_F \pm 2|X - X^*|^2_F |A^T S^T S B^*|^2_F$ (use $\text{tr}(CD) \leq |C||D|$) |
| $\in |SA(X - X^*)|^2_F \pm 2\epsilon|X - X^*|^2_F |B^*|^2_F$ (if we have approx. matrix product) |
| $\in |A(X - X^*)|^2_F \pm \epsilon(|A(X - X^*)|^2_F + 2|X - X^*|^2_F |B^*|)$ (subspace embedding for $A$) |
Affine Embeddings

- We have
  \[ |S(AX - B)|_F^2 - |SB^*|^2_F \in |A(X - X^*)|^2_F \pm \epsilon(|A(X - X^*)|^2_F + 2|X - X^*|_F|B^*|) \]

- Normal equations imply that
  \[ |AX - B|^2_F = |A(X - X^*)|^2_F + |B^*|^2_F \]

- \[ |S(AX - B)|_F^2 - |SB^*|^2_F - (|AX - B|^2_F - |B^*|^2_F) \in \epsilon(|A(X - X^*)|^2_F + 2|X - X^*|_F|B^*|_F) \]
  \[ \in \pm \epsilon(|A(X - X^*)|_F + |B^*|_F)^2 \]
  \[ \in \pm 2\epsilon(|A(X - X^*)|_F^2 + |B^*|_F^2) \]
  \[ = \pm 2\epsilon|AX - B|^2_F \]

- \[ |SB^*|^2_F = (1 \pm \epsilon)|B^*|^2_F \] (this holds with constant probability)
Affine Embeddings

- Know: $|S(AX - B)|^2_F - |SB^*|^2_F - (|AX - B|^2_F - |B^*|^2_F) \in \pm 2\epsilon|AX - B|^2_F$
- Know: $|SB^*|^2_F = (1 \pm \epsilon)|B^*|^2_F$

- $|S(AX - B)|^2_F = (1 \pm 2\epsilon)|AX - B|^2_F + \epsilon|B^*|^2_F$
  
  $= (1 \pm 3\epsilon)|AX - B|^2_F$

- Completes proof of affine embedding!
Affine Embeddings: Missing Proofs

- **Claim:** $|A + B|^2_F = |A|^2_F + |B|^2_F + 2\text{Tr}(A^TB)$

- **Proof:**
  
  $$|A + B|^2_F = \sum_i |A_i + B_i|^2$$

  $$= \sum_i |A_i|^2 + \sum_i |B_i|^2 + 2\langle A_i, B_i \rangle$$

  $$= |A|^2_F + |B|^2_F + 2\text{Tr}(A^TB)$$
Affine Embeddings: Missing Proofs

- Claim: \( \text{Tr}(AB) \leq |A|_F |B|_F \)

- Proof: \( \text{Tr}(AB) = \sum_i \langle A^i, B_i \rangle \) for rows \( A^i \) and columns \( B_i \)

\[
\leq \sum_i |A^i|_2 |B_i|_2 \text{ by Cauchy-Schwarz for each } i
\]

\[
\leq \left( \sum_i |A^i|_2 \right)^{\frac{1}{2}} \left( \sum_i |B_i|_2 \right)^{\frac{1}{2}} \text{ another Cauchy-Schwarz}
\]

\[
= |A|_F |B|_F
\]
Affine Embeddings: Homework Proof

- **Claim**: $|SB^*|_F^2 = (1 \pm \epsilon)|B^*|_F^2$ with constant probability if CountSketch matrix $S$ has $k = O\left(\frac{1}{\epsilon^2}\right)$ rows

- **Proof**:
  - $|SB^*|_F^2 = \sum_i |SB_i^*|_2^2$
  - By our analysis for CountSketch and linearity of expectation,
    $E[|SB^*|_F^2] = \sum_i E[|SB_i^*|_2^2] = |B^*|_F^2$
  - $E[|SB^*|_F^4] = \sum_{i,j} E[|SB_i^*|_2^2 |SB_j^*|_2^2]$
  - By our CountSketch analysis, $E[|SB_i^*|_2^4] \leq |B_i^*|_2^4 \left(1 + \frac{2}{k}\right)$
  - For cross terms see Lemma 40 in [CW13]
Low rank approximation

- A is an n x d matrix
  - Think of n points in $\mathbb{R}^d$

- E.g., A is a customer-product matrix
  - $A_{i,j} =$ how many times customer i purchased item j

- A is typically well-approximated by low rank matrix
  - E.g., high rank because of noise

- **Goal:** find a low rank matrix approximating A
  - Easy to store, data more interpretable
What is a good low rank approximation?

**Singular Value Decomposition (SVD)**

Any matrix $A = U \cdot \Sigma \cdot V$

- $U$ has orthonormal columns
- $\Sigma$ is diagonal with non-increasing positive entries down the diagonal
- $V$ has orthonormal rows

- Rank-$k$ approximation: $A_k = U_k \cdot \Sigma_k \cdot V_k$
  - rows of $V_k$ are the top $k$ principal components

\[
\begin{pmatrix}
A \\
\end{pmatrix} = 
\begin{pmatrix}
U_k \\
\end{pmatrix}
\begin{pmatrix}
\Sigma_k \\
\end{pmatrix}
\begin{pmatrix}
V_k \\
\end{pmatrix} + 
\begin{pmatrix}
E \\
\end{pmatrix}
\]
What is a good low rank approximation?

\[ A_k = \arg\min_{\text{rank } k \text{ matrices } B} |A-B|_F \]

\( |C|_F = (\Sigma_{i,j} C_{i,j}^2)^{1/2} \)

Computing \( A_k \) exactly is expensive

\[
\begin{pmatrix}
A
\end{pmatrix} = \begin{pmatrix}
U_k
\end{pmatrix} \begin{pmatrix}
\Sigma_k
\end{pmatrix} \begin{pmatrix}
V_k
\end{pmatrix} + \begin{pmatrix}
E
\end{pmatrix}
\]
Low rank approximation

- **Goal:** output a rank $k$ matrix $A'$, so that
  \[ |A - A'|_F \leq (1 + \varepsilon) |A - A_k|_F \]

- Can do this in $\text{nnz}(A) + (n + d)\text{poly}(k/\varepsilon)$ time [S,CW]
  - $\text{nnz}(A)$ is number of non-zero entries of $A$
Solution to low-rank approximation [S]

- Given $n \times d$ input matrix $A$
- Compute $S \cdot A$ using a random matrix $S$ with $k/\varepsilon \ll n$ rows. $S \cdot A$ takes random linear combinations of rows of $A$

- Project rows of $A$ onto $S \cdot A$, then find best rank-$k$ approximation to points inside of $S \cdot A$. 
What is the matrix $S$?

- $S$ can be a $k/\varepsilon \times n$ matrix of i.i.d. normal random variables

- $[S]$ $S$ can be a $k/\varepsilon \times n$ Fast Johnson Lindenstrauss Matrix
  - Uses Fast Fourier Transform

- $[CW]$ $S$ can be a $\text{poly}(k/\varepsilon) \times n$ CountSketch matrix

\[
\begin{bmatrix}
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 1 & 0 & -1 & 0 \\
0 & -1 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]

$S \cdot A$ can be computed in $\text{nnz}(A)$ time
Why do these Matrices Work?

- Consider the regression problem $\min_{X} |A_kX - A|_F$

- Let $S$ be an affine embedding

- Then $|SA_kX - SA|_F = (1 \pm \epsilon)|A_kX - A|_F$ for all $X$

- By normal equations, $\arg\min_{X} |SA_kX - SA|_F = (SA_k)^{-1}SA$

- So, $|A_k(_SA_k)^{-1}SA - A|_F \leq (1 + \epsilon)|A_k - A|_F$

- But $A_k(_SA_k)^{-1}SA$ is a rank-$k$ matrix in the row span of $SA$!

- Let’s formalize why the algorithm works now…
Why do these Matrices Work?

- \[ \min_{\text{rank-k } X} |XSA - A|_F^2 \leq |A_k (SA_k)^{-1} SA - A|_F^2 \leq (1 + \epsilon)|A - A_k|_F^2 \]

- By the normal equations,
  \[ |XSA - A|_F^2 = |XSA - A(SA)^{-1} SA|_F^2 + |A(SA)^{-1} SA - A|_F^2 \]

- Hence,
  \[ \min_{\text{rank-k } X} |XSA - A|_F^2 = |A(SA)^{-1} SA - A|_F^2 + \min_{\text{rank-k } X} |XSA - A(SA)^{-1} SA|_F^2 \]

- Can write \( SA = U \Sigma V^T \) in its SVD

- Then, \[ \min_{\text{rank-k } X} |XSA - A(SA)^{-1} SA|_F^2 = \min_{\text{rank-k } X} |XU \Sigma - A(SA)^{-1} U \Sigma|_F^2 \]
  \[ = \min_{\text{rank-k } Y} |Y - A(SA)^{-1} U \Sigma|_F^2 \]

- Hence, we can just compute the SVD of \( A(SA)^{-1} U \Sigma \)

- But how do we compute \( A(SA)^{-1} U \Sigma \) quickly?
Caveat: projecting the points onto SA is slow

- **Current algorithm:**
  1. Compute $S^*A$
  2. Project each of the rows onto $S^*A$
  3. Find best rank-$k$ approximation of projected points inside of rowspace of $S^*A$

- **Bottleneck is step 2**

- **[CW] Approximate the projection**
  - Fast algorithm for approximate regression
    $\min_{\text{rank}-k X} \|X(SA)-RAR\|_F^2$
    Can solve with affine embeddings

- Want $\text{nnz}(A) + (n+d)\text{poly}(k/\epsilon)$ time
Using Affine Embeddings

- We know we can just output \( \arg\min_{\text{rank} - k} \| \text{XSA} - A \|_F^2 \)

- Choose an affine embedding \( R \):
  \[
  \| \text{XSAR} - AR \|_F^2 = (1 \pm \epsilon) \| \text{XSA} - A \|_F^2 \text{ for all } X
  \]

- Note: we can compute \( AR \) and \( SAR \) in \( \text{nnz}(A) \) time

- Can just solve \( \min_{\text{rank} - k} \| \text{XSAR} - AR \|_F^2 \)

- \( \min_{\text{rank} - k} \| \text{XSAR} - AR \|_F^2 = \| \text{AR} \text{(SAR)}^{-1} (\text{SAR}) - AR \|_F^2 + \min_{\text{rank} - k} \| \text{XSAR} - \text{AR(SAR)}^{-1} (\text{SAR}) \|_F^2 \)

- Compute \( \min_{\text{rank} - k} \| Y - \text{AR(SAR)}^{-1} (\text{SAR}) \|_F^2 \) using SVD which is \( (n + d) \text{poly} \left( \frac{k}{\epsilon} \right) \) time

- Necessarily, \( Y = XSAR \) for some \( X \). Output \( Y \text{(SAR)}^{-1} \text{SA} \) in factored form. We’re done!
Low Rank Approximation Summary

1. Compute SA

2. Compute SAR and AR

3. Compute \( \min_{\text{rank} - k} |Y - AR(SAR)^{-}(SAR)|_F^2 \) using SVD

4. Output \( Y(SAR)^{-}SA \) in factored form

Overall time: \( \text{nnz}(A) + (n+d)\text{poly}(k/\epsilon) \)
Course Outline

- Subspace embeddings and least squares regression
  - Gaussian matrices
  - Subsampled Randomized Hadamard Transform
  - CountSketch
- Affine embeddings
  - Application to low rank approximation
- High precision regression
- Leverage score sampling
- Distributed low rank approximation
- L1 Regression
- M-Estimator regression
High Precision Regression

- **Goal**: output $x'$ for which $|Ax'-b|_2 \leq (1+\varepsilon) \min_x |Ax-b|_2$ with high probability

- Our algorithms all have running time $\text{poly}(d/\varepsilon)$

- **Goal**: Sometimes we want running time $\text{poly}(d) \cdot \log(1/\varepsilon)$

- Want to make $A$ well-conditioned
  - $\kappa(A) = \sup_{|x|_2=1} |Ax|_2 / \inf_{|x|_2=1} |Ax|_2$

- Lots of algorithms’ time complexity depends on $\kappa(A)$

- Use sketching to reduce $\kappa(A)$ to $O(1)$!
Small QR Decomposition

- Let $S$ be a $(1 + \epsilon_0)$- subspace embedding for $A$
- Compute $SA$
- Compute QR-factorization, $SA = QR^{-1}$
- Claim: $\kappa(AR) = \frac{(1+\epsilon_0)}{1-\epsilon_0}$
- For all unit $x$, $(1 - \epsilon_0)|ARx|_2 \leq |SARx|_2 = 1$
- For all unit $x$, $(1 + \epsilon_0)|ARx|_2 \geq |SARx|_2 = 1$
- So $\kappa(AR) = \sup_{|x|_2=1} |ARx|_2 / \inf_{|x|_2=1} |ARx|_2 \leq \frac{1+\epsilon_0}{1-\epsilon_0}$
Finding a Constant Factor Solution

- Let $S$ be a $1 + \epsilon_0$ - subspace embedding for $AR$
- Solve $x_0 = \arg\min_x |SARx - Sb|_2$
- Time to compute $R$ and $x_0$ is $\text{nnz}(A) + \text{poly}(d)$ for constant $\epsilon_0$
- $x_{m+1} \leftarrow x_m + R^T A^T (b - AR x_m)$
- $AR(x_{m+1} - x^*) = AR(x_m + R^T A^T (b - AR x_m) - x^*)$  
  $\quad = (AR - AR R^T A^T AR)(x_m - x^*)$  
  $\quad = U(\Sigma - \Sigma^3)V^T(x_m - x^*)$,  
  where $AR = U \Sigma V^T$ is the SVD of $AR$
- $|AR(x_{m+1} - x^*)|_2 = |(\Sigma - \Sigma^3)V^T(x_m - x^*)|_2 = O(\epsilon_0)|AR(x_m - x^*)|_2$  
- $|ARx_m - b|_2^2 = |AR(x_m - x^*)|_2^2 + |ARx^* - b|_2^2$
Course Outline

- Subspace embeddings and least squares regression
  - Gaussian matrices
  - Subsampled Randomized Hadamard Transform
  - CountSketch
- Affine embeddings
  - Application to low rank approximation
- High precision regression
- Leverage score sampling
- Distributed low rank approximation
- M-Estimator regression
This is another subspace embedding, but it is based on sampling!
- If $A$ has sparse rows, then $SA$ has sparse rows!

Let $A = U \Sigma V^T$ be an $n \times d$ matrix with rank $d$, written in its SVD

Define the $i$-th leverage score $\ell(i)$ of $A$ to be $|U_{i,*}|^2$

What is $\sum_i \ell(i)$?
- Let $(q_1, \ldots, q_n)$ be a distribution with $q_i \geq \frac{\beta \ell(i)}{d}$, where $\beta$ is a parameter

Define sampling matrix $S = D \cdot \Omega^T$, where $D$ is $k \times k$ and $\Omega$ is $n \times k$
- $\Omega$ is a sampling matrix, and $D$ is a rescaling matrix
- For each column $j$ of $\Omega, D$, independently, and with replacement, pick a row index $i$ in $[n]$ with probability $q_i$, and set $\Omega_{i,j} = 1$ and $D_{i,j} = \frac{1}{(q_i k)^5}$
Leverage Score Sampling

- Note: leverage scores do not depend on choice of orthonormal basis $U$ for columns of $A$

- Indeed, let $U$ and $U'$ be two such orthonormal bases

- Claim: $|e_i U|^2 = |e_i U'|^2$ for all $i$

- Proof: Since both $U$ and $U'$ have column space equal to that of $A$, we have $U = U'Z$ for change of basis matrix $Z$

- Since $U$ and $U'$ each have orthonormal columns, $Z$ is a rotation matrix (orthonormal rows and columns)

- Then $|e_i U|^2 = |e_i U'Z|^2 = |e_i U'|^2$
Leverage Score Sampling gives a Subspace Embedding

- Want to show for $S = D \cdot \Omega^T$, that $|SAx|_2^2 = (1 \pm \epsilon)|Ax|_2^2$ for all $x$

- Writing $A = U \Sigma V^T$ in its SVD, this is equivalent to showing $|SUy|_2^2 = (1 \pm \epsilon)|Uy|_2^2 = (1 \pm \epsilon)|y|_2^2$ for all $y$

- As usual, we can just show with high probability, $|U^T S^T S U - I|_2 \leq \epsilon$

- How can we analyze $U^T S^T S U$?

- (Matrix Chernoff) Let $X_1, ..., X_k$ be independent copies of a symmetric random matrix $X \in \mathbb{R}^{d \times d}$ with $E[X] = 0$, $|X|_2 \leq \gamma$, and $E[|X^T X|_2] \leq \sigma^2$. Let $W = \frac{1}{k} \sum_{j \in [k]} X_j$. For any $\epsilon > 0$,

$$\Pr[|W|_2 > \epsilon] \leq 2d \cdot e^{-k\epsilon^2/(\sigma^2 + \frac{\gamma^2}{3})}$$

(here $|W|_2 = \sup_{|x|_2} \frac{|Wx|_2}{|x|_2}$. Since $W$ is symmetric, $|W|_2 = \sup_{|x|_2=1} x^T W x$.)
Leverage Score Sampling gives a Subspace Embedding

- Let $i(j)$ denote the index of the row of $U$ sampled in the $j$-th trial
- Let $X_j = I_d - \frac{U_{i(j)}^T U_{i(j)}}{q_{i(j)}}$, where $U_{i(j)}$ is the $j$-th sampled row of $U$
- The $X_j$ are independent copies of a symmetric matrix random variable
- $E[X_j] = I_d - \sum_i q_i \left( \frac{U_i^T U_i}{q_i} \right) = I_d - I_d = 0_d$
- $|X_j|_2 \leq |I_d|_2 + \frac{|U_{i(j)}^T U_{i(j)}|_2}{q_{i(j)}} \leq 1 + \max_i \frac{|U_i|^2}{q_i} \leq 1 + \frac{d}{\beta}$
- $E[X^T X] = I_d - 2E \left[ \frac{U_{i(j)}^T U_{i(j)}}{q_{i(j)}} \right] + E \left[ \frac{U_{i(j)}^T U_{i(j)} U_{i(j)}^T U_{i(j)}}{q_{i(j)}^2} \right]$
  \[= \sum_i \frac{U_i^T U_i U_i^T U_i}{q(i)} - I_d \leq \left( \frac{d}{\beta} \right) \sum_i U_i^T U_i - I_d \leq \left( \frac{d}{\beta} - 1 \right) I_d, \]
  where $A \leq B$ means $x^T Ax \leq x^T Bx$ for all $x$
- Hence, $|E[X^T X]|_2 \leq \frac{d}{\beta} - 1$
Applying the Matrix Chernoff Bound

- (Matrix Chernoff) Let $X_1, ..., X_k$ be independent copies of a symmetric random matrix $X \in \mathbb{R}^{d \times d}$ with $E[X] = 0$, $|X|_2 \leq \gamma$, and $|E[X^T X]|_2 \leq \sigma^2$. Let $W = \frac{1}{k} \sum_{j \in [k]} X_j$. For any $\epsilon > 0$,
  \[ \Pr[|W|_2 > \epsilon] \leq 2d \cdot e^{-k \epsilon^2 / (\sigma^2 + \frac{\gamma \epsilon}{3})} \]
  (here $|W|_2 = \sup_{|x|_2} \frac{|W x|_2}{|x|_2}$. Since $W$ is symmetric, $|W|_2 = \sup_{|x|_2=1} x^T W x$.)

- $\gamma = 1 + \frac{d}{\beta}$ and $\sigma^2 = \frac{d}{\beta} - 1$

- $X_j = I_d - \frac{U_{i(j)}^T U_{i(j)}}{q_{i(j)}}$, and recall how we generated $S = D \cdot \Omega^T$: For each column $j$ of $\Omega, D$, independently, and with replacement, pick a row index $i$ in $[n]$ with probability $q_i$, and set $\Omega_{i,j} = 1$ and $D_{i,j} = 1/(q_i k)$.\footnote{Implies $W = I_d - U^T S^T S U$}
  
  - $\Pr \left[ |I_d - U^T S^T S U|_2 > \epsilon \right] \leq 2d \cdot e^{-k \epsilon^2 \Theta(\frac{\beta}{d})}$. Set $k = \Theta\left(\frac{d \log d}{\beta \epsilon^2}\right)$ and we’re done.
Fast Computation of Leverage Scores

- Naively, need to do an SVD to compute leverage scores
- Suppose we compute \( SA \) for a subspace embedding \( S \)
- Let \( SA = QR^{-1} \) be such that \( Q \) has orthonormal columns
- Set \( \ell_i' = |e_iAR|^2 \)
- Since \( AR \) has the same column span of \( A \), \( AR = UT^{-1} \)
  - \( (1 - \epsilon)|ARx|_2 \leq |SARx|_2 = |x|_2 \)
  - \( (1 + \epsilon)|ARx|_2 \geq |SARx|_2 = |x|_2 \)
  - \( (1 \pm O(\epsilon))|x|_2 = |ARx|_2 = |UT^{-1}x|_2 = |T^{-1}x|_2 \),
- \( \ell_i = |e_iART|^2 = (1 \pm O(\epsilon))|e_iAR|^2 = (1 \pm O(\epsilon))\ell_i' \)
- But how do we compute \( AR \)? We want \( \text{nnz}(A) \) time
Fast Computation of Leverage Scores

- $\ell_i = (1 \pm O(\epsilon))\ell_i'$
  - Suffices to set this $\epsilon$ to be a constant

- Set $\ell_i' = |e_iAR|^2$
  - This takes too long

- Let $G$ be a $d \times O(\log n)$ matrix of i.i.d. normal random variables
  - For any vector $z$, $\Pr[|zG|^2 = (1 \pm \frac{1}{2})|z|^2] \geq 1 - \frac{1}{n^2}$

- Instead set $\ell_i' = |e_iARG|^2$.
  - Can compute in $(\text{nnz}(A) + d^2)\log n$ time

- Can solve regression in $\text{nnz}(A) \log n + \text{poly}(d(\log n)/\epsilon)$ time
Course Outline

- Subspace embeddings and least squares regression
  - Gaussian matrices
  - Subsampled Randomized Hadamard Transform
  - CountSketch
- Affine embeddings
  - Application to low rank approximation
- High precision regression
- Leverage score sampling
- Distributed low rank approximation
- L1 Regression
- M-Estimator regression
Distributed low rank approximation

- We have fast algorithms for low rank approximation, but can they be made to work in a distributed setting?

- Matrix $A$ distributed among $s$ servers

- For $t = 1, \ldots, s$, we get a customer-product matrix from the $t$-th shop stored in server $t$. Server $t$’s matrix $= A^t$

- Customer-product matrix $A = A^1 + A^2 + \ldots + A^s$
  - Model is called the arbitrary partition model

- More general than the row-partition model in which each customer shops in only one shop
The Communication Model

- Each player talks only to a Coordinator via 2-way communication.
- Can simulate arbitrary point-to-point communication up to factor of 2 (and an additive $O(\log s)$ factor per message).
Communication cost of low rank approximation

- **Input:** n x d matrix A stored on s servers
  - Server t has n x d matrix $A^t$
  - $A = A^1 + A^2 + \ldots + A^s$
  - Assume entries of $A^t$ are $O(\log(nd))$-bit integers

- **Output:** Each server outputs the same k-dimensional space W
  - $C = A^1P_W + A^2P_W + \ldots + A^sP_W$, where $P_W$ is the projection onto W
  - $|A-C|_F \leq (1+\varepsilon)|A-A_{k}\|_F$
  - Application: k-means clustering

- **Resources:** Minimize total communication and computation. Also want $O(1)$ rounds and input sparsity time
Work on Distributed Low Rank Approximation

- [FSS]: First protocol for the row-partition model.
  - $O(sdk/\varepsilon)$ real numbers of communication
  - Don’t analyze bit complexity (can be large)
  - SVD Running time, see also [BKLW]

- [KVW]: $O(skd/\varepsilon)$ communication in arbitrary partition model

- [BWZ]: $O(skd) + \text{poly}(sk/\varepsilon)$ words of communication in arbitrary partition model. Input sparsity time
  - Matching $\Omega(skd)$ words of communication lower bound

- Variants: kernel low rank approximation [BLSWX], low rank approximation of an implicit matrix [WZ], sparsity [BWZ]
Outline of Distributed Protocols

- [FSS] protocol
- [KVW] protocol
- [BWZ] protocol
Constructing a Coreset [FSS]

- Let $A = U \Sigma V^T$ be its SVD

- Let $m = k + k/\epsilon$

- Let $\Sigma_m$ agree with $\Sigma$ on the first $m$ diagonal entries, and be 0 otherwise

- Claim: For all projection matrices $Y = I - X$ onto (d-k)-dimensional subspaces,

$$|\Sigma_m V^T Y|_F^2 = (1 \pm \epsilon) |AY|_F^2 + c,$$

where $c = |A - A_m|_F^2$ does not depend on $Y$

- We can think of $S$ as $U^T_m$ so that $SA = U^T_m U \Sigma V^T = \Sigma_m V^T$ is a sketch
Constructing a Coreset

Claim: For all projection matrices \( Y=I-X \) onto (n-k)-dimensional subspaces,

\[
\|\Sigma_mV^TY\|_F^2 + c = (1 \pm \epsilon)|AY|_F^2,
\]

where \( c = |A - A_m|_F^2 \) does not depend on \( Y \)

Proof: \( |AY|_F^2 = |U\Sigma_mV^TY|_F^2 + |U(\Sigma - \Sigma_m)V^TY|_F^2 \)

\[
\leq \|\Sigma_mV^TY\|_F^2 + |A - A_m|_F^2 = \|\Sigma_mV^TY\|_F^2 + c
\]

Also, \( \|\Sigma_mV^TY\|_F^2 + |A - A_m|_F^2 - |AY|_F^2 \)

\[
= \|\Sigma_mV^T\|_F^2 - \|\Sigma_mV^TX\|_F^2 + |A - A_m|_F^2 - |A|_F^2 + |AX|_F^2
\]

\[
= |AX|_F^2 - \|\Sigma_mV^TX\|_F^2
\]

\[
= \|U(\Sigma - \Sigma_m)V^TX\|_F^2
\]

\[
\leq \|U(\Sigma - \Sigma_m)V\|_F^2 \cdot |X|_F^2
\]

\[
\leq \sigma^2_{m+1} k \leq \epsilon \sigma^2_{m+1} (m - k) \leq \epsilon \sum_{i \in \{k+1, \ldots, m+1\}} \sigma^2_i \leq \epsilon |A - A_k|_F^2
\]

89
Unions of Coresets

- Suppose we have matrices $A^1, \ldots, A^s$ and construct $\Sigma^1_m V^{T,1}, \Sigma^2_m V^{T,2}, \ldots, \Sigma^s_m V^{T,s}$ as in the previous slide, together with $c_1, \ldots, c_s$

- Then $\sum_i |\Sigma^i_m V^{T,i}Y|_F^2 + c_i = (1 \pm \epsilon)|AY|_F^2$, where $A$ is the matrix formed by concatenating the rows of $A^1, \ldots, A^s$

- Let $B$ be the matrix obtained by concatenating the rows of $\Sigma^1_m V^{T,1}, \Sigma^2_m V^{T,2}, \ldots, \Sigma^s_m V^{T,s}$

- Suppose we compute $B = U \Sigma V^T$ and compute $\Sigma_m V^T$ and $|B - B_m|_F^2$

- Then $|\Sigma_m V^T Y|_F^2 + c + \sum_i c_i = (1 \pm \epsilon)|BY|_F^2 + \sum_i c_i = (1 \pm O(\epsilon))|AY|_F^2$

- So $\Sigma_m V^T$ and the constant $c + \sum_i c_i$ are a coreset for $A$
[FSS] Row-Partition Protocol

- Server $t$ sends the top $k/\varepsilon + k$ principal components of $P^t$, scaled by the top $k/\varepsilon + k$ singular values $\Sigma^t$, together with $c^t$.

- Coordinator returns top $k$ principal components of $[\Sigma^1 V^1; \Sigma^2 V^2; \ldots; \Sigma^s V^s]$. 

$P^1 \in \mathbb{R}^{n_1 \times d}$

$P^2 \in \mathbb{R}^{n_2 \times d}$

$P^s \in \mathbb{R}^{n_s \times d}$

Coordinator
[FSS] Row-Partition Protocol

Problems:
1. $\text{sdk/}\varepsilon$ real numbers of communication
2. bit complexity can be large
3. running time for SVDs [BLKW]
4. doesn’t work in arbitrary partition model

This is an SVD-based protocol. Maybe our random matrix techniques can improve communication just like they improved computation?

[KVW] protocol will handle 2, 3, and 4

- Inspired by the sketching algorithm presented earlier

- Let $S$ be one of the $k/\varepsilon \times n$ random matrices discussed
  - $S$ can be generated pseudorandomly from small seed
  - Coordinator sends small seed for $S$ to all servers

- Server $t$ computes $SA^t$ and sends it to Coordinator

- Coordinator sends $\sum_{t=1}^{s} SA^t = SA$ to all servers

- There is a good $k$-dimensional subspace inside of $SA$. If we knew it, $t$-th server could output projection of $A^t$ onto it
[KVW] Arbitrary Partition Model Protocol

Problems:

- Can’t output projection of $A^t$ onto SA since the rank is too large

- Could communicate this projection to the coordinator who could find a k-dimensional space, but communication depends on $n$
Fix:

- Instead of projecting $A$ onto $SA$, recall we can solve
  \[ \min_{\text{rank-}k X} \left\| A(SA)^T XSA - A \right\|_F^2 \]
- Let $T_1, T_2$ be affine embeddings, solve
  \[ \min_{\text{rank-}k X} \left\| T_1 A(SA)^T XSAT_2 - T_1 AT_2 \right\|_F^2 \]
  (optimization problem is small and has a closed form solution)
- Everyone can then compute $XSA$ and then output $k$ directions
[KVW] protocol

- Phase 1:
  - Learn the row space of SA

\[
\text{cost} \leq (1+\varepsilon)|A-A_k|_F
\]
[KVW] protocol

- Phase 2:
  - Find an approximately optimal space $W$ inside of $SA$

$$\text{cost} \leq (1+\epsilon)^2|A-A_k|_F$$
Main Problem: communication is $O(\text{skd}/\varepsilon) + \text{poly}(sk/\varepsilon)$

We want $O(\text{skd}) + \text{poly}(sk/\varepsilon)$ communication!

Idea: use projection-cost preserving sketches [CEMMP]

Let $A$ be an $n \times d$ matrix

If $S$ is a random $k/\varepsilon^2 \times n$ matrix, then there is a constant $c \geq 0$ so that for all $k$-dimensional projection matrices $P$:

$$|SA(I - P)|_F + c = (1 \pm \varepsilon)|A(I - P)|_F$$
[BWZ] Protocol

Intuitively, $U$ looks like top $k$ left singular vectors of $SA$.

- Let $S$ be a $k/\varepsilon^2 \times n$ projection-cost preserving sketch.
- Let $T$ be a $d \times k/\varepsilon^2$ projection-cost preserving sketch.
- Server $t$ sends $SA^tT$ to Coordinator.

Coordinator sends back $SAT = \sum_t SA^tT$ to servers.

- Each server computes $k/\varepsilon^2 \times k$ matrix $U$ of top $k$ left singular vectors of $SAT$.
- $U^{TSA}$ looks like top $k$ right singular vectors of $SA$.

Server $t$ sends $U^{TSA^t}$ to Coordinator.

Coordinator returns the space $U^{TSA} = \sum_t U^{TSA^t}$ to output.

Top $k$ right singular vectors of $SA$ work because $S$ is a projection-cost preserving sketch!
[BWZ] Analysis

- Let $W$ be the row span of $U^T S A$, and $P$ be the projection onto $W$

- Want to show $|A - AP|_F \leq (1 + \epsilon)|A - A_k|_F$

- Since $T$ is a projection-cost preserving sketch,

\[(*) \quad |SA - SAP|_F \leq |SA - UU^T S A|_F + c_1 \leq (1 + \epsilon)|SA - [SA]_k|_F\]

- Since $S$ is a projection-cost preserving sketch, there is a scalar $c > 0$, so that for all $k$-dimensional projection matrices $Q$,

\[|SA - SAQ|_F + c = (1 \pm \epsilon)|A - AQ|_F\]

- Add $c$ to both sides of $(*)$ to conclude $|A - AP|_F \leq (1 + \epsilon)|A - A_k|_F$
Conclusions for Distributed Low Rank Approximation

- [BWZ] Optimal $O(sdk) + \text{poly}(sk/\varepsilon)$ communication protocol for low rank approximation in arbitrary partition model
  - Handle bit complexity by adding noise
  - Input sparsity time
  - 2 rounds, which is optimal [W]
  - Optimal data stream algorithms improves [CW, L, GP]

- Communication of other optimization problems?
  - Computing the rank of an $n \times n$ matrix over the reals
  - Linear Programming
  - Graph problems: Matching
  - etc.
Course Outline

- Subspace embeddings and least squares regression
  - Gaussian matrices
  - Subsampled Randomized Hadamard Transform
  - CountSketch
- Affine embeddings
  - Application to low rank approximation
- High precision regression
- Leverage score sampling
- Distributed low rank approximation
- L1 Regression
- M-Estimator Regression
Robust Regression

Method of least absolute deviation ($l_1$-regression)

- Find $x^*$ that minimizes $|Ax - b|_1 = \sum |b_i - \langle A_{i*}, x \rangle|$

- Cost is less sensitive to outliers than least squares

- Can solve via linear programming
Solving $l_1$-regression via Linear Programming

- Minimize $(1,\ldots,1) \cdot (\alpha^+ + \alpha^-)$
- Subject to:
  
  \[ A \cdot x + \alpha^+ - \alpha^- = b \]
  
  \[ \alpha^+, \alpha^- \geq 0 \]

- Generic linear programming gives poly(nd) time

- Want much faster time using sketching!
Well-Conditioned Bases

- For an \( n \times d \) matrix \( A \), can choose an \( n \times d \) matrix \( U \) with orthonormal columns for which \( A = UW \), and \( |Ux|_2 = |x|_2 \) for all \( x \)

- Can we find a \( U \) for which \( A = UW \) and \( |Ux|_1 \approx |x|_1 \) for all \( x \)?

- Let \( A = QW \) where \( Q \) has full column rank, and define \( |z|_{Q,1} = |Qz|_1 \)
  - \( |z|_{Q,1} \) is a norm

- Let \( C = \{z \in \mathbb{R}^d : |z|_{Q,1} \leq 1\} \) be the unit ball of \( |.|_{Q,1} \)

- \( C \) is a convex set which is symmetric about the origin
  - Lowner-John Theorem: can find an ellipsoid \( E \) such that: \( E \subseteq C \subseteq \sqrt{d}E \), where \( E = \{z \in \mathbb{R}^d : z^TFz \leq 1\} \)
    - \( (z^TFz)^{\frac{5}{2}} \leq |z|_{Q,1} \leq \sqrt{d}(z^TFz)^{\frac{5}{2}} \)
    - \( F = GG^T \) since \( F \) defines an ellipsoid

- Define \( U = QG^{-1} \)
Well-Conditioned Bases

- Recall $U = QG^{-1}$ where
  \[(z^T Fz)^{\frac{5}{2}} \leq |z|_{Q,1} \leq \sqrt{d}(z^T Fz)^{\frac{5}{2}} \text{ and } F = GG^T\]

- $|Ux|_1 = |QG^{-1}x|_1 = |Qz|_1 = |z|_{Q,1}$ where $z = G^{-1}x$

- $z^T Fz = (x^T(G^{-1})^T G G (G^{-1})x) = x^T x = |x|_2^2$

- So $|x|_2 \leq |Ux|_1 \leq \sqrt{d}|x|_2$

- So $\frac{|x|_1}{\sqrt{d}} \leq |x|_2 \leq |Ux|_1 \leq \sqrt{d}|x|_2 \leq \sqrt{d}|x|_1$
Net for $\ell_1$ – Ball

- Consider the unit $\ell_1$-ball $B = \{x \in \mathbb{R}^d : |x|_1 = 1\}$
- Subset $N$ is a $\gamma$-net if for all $x \in B$, there is a $y \in N$, such that $|x - y|_1 \leq \gamma$
- Greedy construction of $N$
  - While there is a point $x \in B$ of distance larger than $\gamma$ from every point in $N$, include $x$ in $N$
- The $\ell_1$-ball of radius $\gamma/2$ around every point in $N$ is contained in the $\ell_1$-ball of radius $1 + \gamma/2$ around $0^d$
- Further, all such ball are disjoint
- Ratio of volume of $d$-dimensional similar polytopes of radius $1 + \gamma/2$ to radius $\gamma/2$ is $(1 + \gamma/2)^d/(\gamma/2)^d$, so $|N| \leq (1 + \gamma/2)^d/(\gamma/2)^d$
Net for $\ell_1$ – Subspace

- Let $A = UW$ for a well-conditioned basis $U$
  - $|x|_1 \leq |Ux|_1 \leq d|x|_1$ for all $x$

- Let $N$ be a $(\gamma/d)$–net for the unit $\ell_1$-ball $B$

- Let $M = \{Ux \mid x \in N\}$, so $|M| \leq (1 + \gamma/(2d))^d/((\gamma/(2d))^d$

- Claim: For every $x$ in $B$, there is a $y$ in $M$ for which $|Ax - y|_1 \leq \gamma$

- Proof: Let $x'$ in $B$ be such that $|x - x'|_1 \leq \gamma/d$
  Then $|Ax - Ax'|_1 \leq d|x - x'|_1 \leq \gamma$, using the well-conditioned basis property. Set $y = Ax'$

- $|M| \leq \left(\frac{d}{\gamma}\right)^{O(d)}$
Rough Algorithm Overview

1. Compute poly(d)-approximation
2. Compute well-conditioned basis
3. Sample rows from the well-conditioned basis and the residual of the poly(d)-approximation
4. Solve $l_1$-regression on the sample, obtaining vector $x$, and output $x$

Takes $\text{nnz}(A)$ time
Takes $\text{poly}(d/\varepsilon)$ time
Rough Algorithm Overview

\[ \min_{x \in \mathbb{R}^d} |Ax-b|_1 = \min_{x \in \mathbb{R}^d} |Ux - b'|_1 \]

Sample \( \text{poly}(d/\varepsilon) \) rows of \( U \circ b' \) proportional to their \( l_1 \)-norm.

Find \( x' \) such that
\[ |Ax'-b|_1 \leq \text{poly}(d) \min_{x \in \mathbb{R}^d} |Ax-b|_1 \]
Let \( b' = b - Ax' \) be the residual.

Find a basis \( A = UW \) so that for all \( x \in \mathbb{R}^d \),
\[ |x|_1/\text{poly}(d) \leq |Ux|_1 \leq \text{poly}(d) |x|_1 \]
Now generic linear programming is efficient.
Will focus on showing how to quickly compute

1. A poly(d)-approximation

2. A well-conditioned basis
Sketching Theorem

Theorem

- There is a probability space over \((d \log d) \times n\) matrices \(R\) such that for any \(n \times d\) matrix \(A\), with probability at least \(99/100\) we have for all \(x\):
  \[
  |Ax|_1 \leq |RAx|_1 \leq d \log d \cdot |Ax|_1
  \]

Embedding

- is linear
- is independent of \(A\)
- preserves lengths of an infinite number of vectors
Application of Sketching Theorem

Computing a $d(\log d)$-approximation

- Compute $RA$ and $Rb$
- Solve $x' = \arg\min_x |RAx - Rb|_1$
- Main theorem applied to $A \circ b$ implies $x'$ is a $d \log d$ – approximation
- $RA$, $Rb$ have $d \log d$ rows, so can solve $l_1$-regression efficiently
Application of Sketching Theorem

Computing a well-conditioned basis

1. Compute RA

2. Compute W so that RAW is orthonormal (in the $l_2$-sense)

3. Output $U = AW$

$U = AW$ is well-conditioned because

$$|AWx|_1 \leq |RAWx|_1 \leq (d \log d)^{1/2} |RAWx|_2 = (d \log d)^{1/2} |x|_2 \leq (d \log d)^{1/2} |x|_1$$

and

$$|AWx|_1 \geq |RAWx|_1/(d \log d) \geq |RAWx|_2/(d \log d) = |x|_2/(d \log d) \geq |x|_1/(d^{3/2} \log d)$$
Theorem:

There is a probability space over \((d \log d) \times n\) matrices \(R\) such that for any \(n \times d\) matrix \(A\), with probability at least 99/100 we have for all \(x\):

\[
|Ax|_1 \leq |RAx|_1 \leq d \log d \cdot |Ax|_1
\]

A dense \(R\) that works:

The entries of \(R\) are i.i.d. Cauchy random variables, scaled by \(1/(d \log d)\)
Cauchy Random Variables

- pdf(z) = 1/(π(1+z^2)) for z in (-∞, ∞)
- Undefined expectation and infinite variance
- 1-stable:
  - If z_1, z_2, ..., z_n are i.i.d. Cauchy, then for a ∈ R^n,
    \[ a_1 \cdot z_1 + a_2 \cdot z_2 + \ldots + a_n \cdot z_n \sim |a|_1 \cdot z, \] where z is Cauchy
- Can generate as the ratio of two standard normal random variables
Proof of Sketching Theorem

- By 1-stability,
  - For all rows r of R,
    - \(<r, Ax> = |Ax|_1 \cdot Z / (d \log d)\),
      where Z is a Cauchy
  - \(RAx = (|Ax|_1 \cdot Z_1, ..., |Ax|_1 \cdot Z_{d \log d}) / (d \log d)\),
    where \(Z_1, ..., Z_{d \log d}\) are i.i.d. Cauchy

- \(|RAx|_1 = |Ax|_1 \sum_j |Z_j| / (d \log d)\)
  - The \(|Z_j|\) are half-Cauchy

- \(\sum_j |Z_j| = \Omega(d \log d)\) with probability \(1-\exp(-d \log d)\) by Chernoff

- But the \(|Z_j|\) are heavy-tailed…
Proof of Sketching Theorem

- $\sum_j |Z_j|$ is heavy-tailed, so $|RAx|_1 = |Ax|_1 \sum_j |Z_j| / (d \log d)$ may be large

- Each $|Z_j|$ has c.d.f. asymptotic to $1-\Theta(1/z)$ for $z$ in $[0, \infty)$

- There exists a well-conditioned basis of $A$
  - Suppose w.l.o.g. the basis vectors are $A_{*1}, \ldots, A_{*d}$

- $|RA_{*i}|_1 = |A_{*i}|_1 \cdot \sum_j |Z_{i,j}| / (d \log d)$

- Let $E_{i,j}$ be the event that $|Z_{i,j}| \leq d^3$
  - Define $Z'_{i,j} = |Z_{i,j}|$ if $|Z_{i,j}| \leq d^3$, and $Z'_{i,j} = d^3$ otherwise
  - $E[Z_{i,j} \mid E_{i,j}] = E[Z'_{i,j} \mid E_{i,j}] = O(\log d)$

- Let $E$ be the event that for all $i,j$, $E_{i,j}$ occurs
  - $\Pr[E] \geq 1 - \frac{\log d}{d}$
  - What is $E[Z'_{i,j} \mid E]$?
Proof of Sketching Theorem

- What is $E[Z'_{i,j} \mid E]$?

\[
E[Z'_{i,j} \mid E_{i,j}] = E[Z'_{i,j} \mid E_{i,j}, E] \Pr[ E \mid E_{i,j}] + E[Z'_{i,j} \mid E_{i,j}, \neg E] \Pr[ \neg E \mid E_{i,j}] \\
\geq E[Z'_{i,j} \mid E_{i,j}, E] \Pr[ E \mid E_{i,j}] \\
= E[Z'_{i,j} \mid E] \cdot \left( \frac{\Pr[ E_{i,j} \mid E] \Pr[ E]}{\Pr[ E_{i,j}]} \right) \\
\geq E[Z'_{i,j} \mid E] \cdot \left( 1 - \frac{\log d}{d} \right)
\]

- So, $E[Z'_{i,j} \mid E] = O(\log d)$

- $|RA^*_i|_1 = |A^*_i|_1 \cdot \sum_{i,j} |Z_{i,j}| / (d \log d)$

- With constant probability, $\sum_i |RA^*_i|_1 = O(\log d) \sum_i |A^*_i|_1$
Proof of Sketching Theorem

- With constant probability, $\sum_i |RA_*i|_1 = O(\log d) \sum_i |A_*i|_1$

- Recall $A_1^*, ..., A_d^*$ is a well-conditioned basis, and we showed the existence of such a basis earlier.

- We will use the Auerbach basis which always exists:
  - For all $x$, $|x|_\infty \leq |Ax|_1$
  - $\sum_i |A_*i|_1 = d$

- $\sum_i |RA_*i|_1 = O(d \log d)$

- For all $x$, $|RAx|_1 \leq \sum_i |RA_*i x_i| \leq |x|_\infty \sum_i |RA_*i|_1$
  $= |x|_\infty O(d \log d)$
  $= O(d \log d) |Ax|_1$
Where are we?

- Suffices to show for all $x$ with $|x|_1 = 1$, that $|Ax|_1 \leq |RAX|_1 \leq d \log d \cdot |Ax|_1$
- We know
  - (1) there is a $\gamma$-net $M$, with $|M| \leq \left(\frac{d}{\gamma}\right)^{O(d)}$, of the set $\{Ax \text{ such that } |x|_1 = 1\}$
  - (2) for any fixed $x$, $|RAX|_1 \geq |Ax|_1$ with probability $1 - \exp(-d \log d)$
  - (3) for all $x$, $|RAX|_1 = O(d \log d)|Ax|_1$

- Set $\gamma = 1/(d^3 \log d)$ so $|M| \leq d^{O(d)}$
  - By a union bound, for all $y$ in $M$, $|Ry|_1 \geq |y|_1$

- Let $x$ with $|x|_1 = 1$ be arbitrary. Let $y$ in $M$ satisfy $|Ax - y|_1 \leq \gamma = 1/(d^3 \log d)$

  - $|RAX|_1 \geq |Ry|_1 - |R(Ax - y)|_1$
    - $\geq |y|_1 - O(d \log d)|Ax - y|_1$
    - $\geq |y|_1 - O(d \log d)\gamma$
    - $\geq |y|_1 - O\left(\frac{1}{d^2}\right)$
    - $\geq |y|_1/2$ (why?)
Sketching to solve $l_1$-regression [CW, MM]

- Most expensive operation is computing $R^*A$ where $R$ is the matrix of i.i.d. Cauchy random variables.

- All other operations are in the “smaller space”.

- Can speed this up by choosing $R$ as follows:

$$
\begin{bmatrix}
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 1 & 0 & -1 & 0 \\
0 & -1 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\cdot
\begin{bmatrix}
C_1 \\
C_2 \\
C_3 \\
\vdots \\
C_n
\end{bmatrix}
$$
Further sketching improvements [WZ]

- Can show you need a fewer number of sampled rows in later steps if instead choose R as follows

- Instead of diagonal of Cauchy random variables, choose diagonal of reciprocals of exponential random variables

\[
\begin{bmatrix}
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 1 & 0 & -1 & 0 \\
0 & -1 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
1/E_1 \\
1/E_2 \\
1/E_3 \\
\vdots \\
1/E_n
\end{bmatrix}
\]
Course Outline

- Subspace embeddings and least squares regression
  - Gaussian matrices
  - Subsampled Randomized Hadamard Transform
  - CountSketch
- Affine embeddings
  - Application to low rank approximation
- High precision regression
- Leverage score sampling
- Distributed low rank approximation
- L1 Regression
- M-Estimator Regression
Robust Regression Fitness Measures

Example: Method of least absolute deviation ($l_1$-regression)

• Find $x^*$ that minimizes $|Ax-b|_1 = \Sigma |b_i - \langle A_{i*}, x \rangle|$

• Cost is less sensitive to outliers than least squares

• Can solve via linear programming

• Can solve in $\text{nnz}(A) + \text{poly}(d/\epsilon)$ time using sketching

What about the many other fitness measures used in practice?
M-Estimators

- **Measure function**
  - $M: \mathbb{R} \rightarrow \mathbb{R}^0$
  - $M(x) = M(-x), M(0) = 0$
  - $M$ is non-decreasing in $|x|$

- $|y|_M = \sum_{i=1}^{n} M(y_i)$

- Solve $\min_x |Ax-b|_M$

- Least squares and $L_1$-regression are special cases
Huber Loss Function

\[ M(x) = \frac{x^2}{2c} \text{ for } |x| \leq c \]

\[ M(x) = |x| - c/2 \text{ for } |x| > c \]

Enjoys smoothness properties of \( l_2^2 \) and robustness properties of \( l_1 \)
Other Examples

• $L_1$-$L_2$
  \[ M(x) = 2((1+x^2/2)^{1/2} - 1) \]

• Fair estimator
  \[ M(x) = c^2 \left[ \frac{|x|}{c} - \log(1+|x|/c) \right] \]

• Tukey estimator
  \[ M(x) = \begin{cases} 
  c^2/6 (1-[1-(x/c)^2]^3) & \text{if } |x| \leq c \\
  c^2/6 & \text{if } |x| > c 
  \end{cases} \]
Nice $M$-Estimators

- An $M$-Estimator is nice if it has at least linear growth and at most quadratic growth.

- There is $C_M > 0$ so that for all $a, a'$ with $|a| \geq |a'| > 0$,
  $$|a/a'|^2 \geq \frac{M(a)}{M(a')} \geq C_M |a/a'|$$

- Any convex $M$ satisfies the linear lower bound (why?)
  $$M(a') = M\left(\left(\frac{a'}{a}\right) \cdot a + \left(1 - \frac{a'}{a}\right) \cdot 0\right) \leq \left(\frac{a'}{a}\right) M(a) + \left(1 - \frac{a'}{a}\right) M(0) = \left(\frac{a'}{a}\right) M(a)$$

- Any sketchable $M$ satisfies the quadratic upper bound
  - sketchable $\Rightarrow$ there is a distribution on $k \times n$ matrices $S$ for which $|Sx|_M = \Theta(|x|_M)$ with good probability and $k$ is slow-growing function of $n$
Nice M-Estimator Theorem

[Nice M-Estimators] $O(\text{nnz}(A)) + \text{poly}(d \log n)$ time algorithm to output $x'$ so that for any constant $C > 1$, with probability 99%:

$$|Ax' - b|_M \leq C \min_x |Ax - b|_M$$

Remarks:

- For convex nice M-estimators can solve with convex programming, but slow – poly(nd) time
- Our sketch is “universal”
\[ T = \begin{bmatrix}
S^0 \cdot D^0 \\
S^1 \cdot D^1 \\
S^2 \cdot D^2 \\
\vdots \\
S^{\log n} \cdot D^{\log n}
\end{bmatrix} \]

- \( S^i \) are independent CountSketch matrices with poly(d) rows
- \( D^i \) is \( n \times n \) diagonal and uniformly samples a \( 1/(d \log n)^i \) fraction of the \( n \) rows
- The same M-Sketch works for all nice M-estimators!

\[ x' = \arg\min_x |TAx - Tb|_{w,M} \]

- many analyses of this data structure don’t work since they reduce the problem to a non-convex problem

- Sketch used for estimating frequency moments [Indyk, W] and earthmover distance [Verbin, Zhang]
M-Sketch Intuition

• For a given $y = Ax - b$, consider $|Ty|_{w,M} = \sum_i w_i M((Ty)_i)$

• [Contraction] $|Ty|_{w,M} \geq \frac{1}{2} |y|_M$ with probability $1 - \exp(-d \log n)$

• [Dilation] $|Ty|_{w,M} \leq 2 |y|_M$ with probability 99%

• Contraction allows for a net argument (no scale-invariance!)
  – Show that $|y^*_2$ is within a factor $\text{poly}(n)$ of $\min_x |Ax - b|_2$

• Dilation implies the optimal $y^*$ does not dilate much

• Proof: try to estimate contribution to $|y|_M$ at all scales
  – E.g., if $y = (n, 1, 1, \ldots, 1)$ with a total of $n-1$ 1s, then $|y|_1 = n + (n-1)*1$
  – When estimating a given scale, use the fact that smaller stuff cancels each other out in a bucket and gives its 2-norm